Research on Painting Multi-Style Transfer Based on Perceptual Loss

Liucan Zhou[a]
College of Computer Science and Electronic Engineering, Hunan University, Hunan, China

Keywords: Convolutional Neural Network, Deep Learning, Image Style Transfer, VGGNet, Image Style Transformation.

Abstract: The art of various paintings has always been a reflection of people's constant pursuit of aesthetics. Style transfer technology integrates various painting styles with image content, bringing this pursuit into people's daily lives. Moreover, the development of deep learning technology and CNN application, style transfer technology has made significant breakthroughs. This paper mainly is about the method of style transfer based on CNN technology. On this topic, the Visual Geometry Group-19 (VGG-19) model is mainly used. By inputting preprocessed images into the VGG-19 model, it aids in isolating the style and content characteristics of images, which in turn facilitates the optimization of the perceptual loss function, and then uses the extracted perceptual loss function to implement a feedforward network for image transformation. This method can achieve satisfactory results and effectively completes the deep fusion of images with the desired style. This not only enhances the practicality of style transfer technology but also provides a new way for the pursuit of aesthetics.

1 INTRODUCTION

With the rapid development of artificial intelligence technology, people's requirements for image processing are no longer satisfied with simple editing and beautification but are pursuing deeper levels of stylization and personalized expression. The emergence of style transfer technology has just met this demand.

Style transfer refers to rendering the content of one image into the style of another image, such as displaying a realistic skyscraper in the style of Van Gogh's painting. The application range of style transfer technology is extremely wide, such as enhancing the efficiency of artistic creation. Style transfer technology can help artists and designers quickly realize creative concepts, reduce production costs, and improve work efficiency. In movie production, through style transfer technology, real scenes can be quickly transformed into scenes with specific artistic styles, bringing unique visual experiences to the audience.

Based on the review by Liu, and Zhou, this paper refers to the methods of style transfer mentioned by different authors. (Liu, Zhou, 2023) Gatys et al. developed a neural network-based algorithm that can effectively separate the content from the style of an image and recombine them, thus creating new images of good quality (Gatys, Ecker, Bethge, 2016). However, during the process of image style conversion, it is necessary to go through a complete training iteration process. This makes the style transfer process time-consuming and not very effective in optimization issues. Subsequently, Johnson et al. proposed using a pre-trained network to extract image features to calculate the perceptual loss (Johnson, Alahi, & Fei-Fei, 2016). This method also optimizes the loss of feature reconstruction, ensuring the characteristics of the image to a large extent, and then uses the loss value between the generated image and the input image for iteration. This ensures that only one training is required. By using this method, efficiency is also improved while ensuring quality. In addition, the CycleGAN (Cycle Generative Adversarial Network) has also been proposed for style transfer. The Cycle GAN model proposed by ZHU et al. Throughout the training phase, it fine-tunes both the generator and the discriminator. The role of the generator is to produce blended images, whereas the discriminator is responsible for identifying the difference between generated and actual images. Moreover, they

[a] https://orcid.org/0009-0002-2378-1375

Zhou and L.
Research on Painting Multi-Style Transfer Based on Perceptual Loss.
DOI: 10.5220/0013524800004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 401-405
ISBN: 978-989-758-754-2

introduced a cycle consistency loss function, allowing the model to complete the task of image transformation without paired data (Zhu, Park, Isola, et al., 2017). However, when using Cycle GAN to train the dataset, it encounters a longer training cycle, and the completion of model training requires a significant amount of time and computational resources, posing higher requirements for hardware configuration. To solve this problem, SI, WANG proposed using the ModileNetV2-CycleGAN model for image style transfer and introduced multi-scale structural similarity (MS-SSIM) as a penalty term to preserve the quality of style images. Furthermore, it improves the effect of feature learning and the quality of stylized images (Si, Wang, 2024).

This paper mainly builds on the research of the Johnson method and uses the VGG model to extract the loss function to complete the generative network, aiming to achieve high-quality style transfer results.

2 RESEARCH DESIGN AND METHODS

2.1 Design of Style Transfer Network

As depicted in Figure 1, in line with the studies conducted by Zhang on the design of style transfer networks, the architecture for image style transfer can be broadly categorized into two components: the transformation network and the extracted loss network (Zhang, Huang, 2023).

Transformation Network: Using the torch library in Python, the transformation network can be regarded as a static image generator that generates target-style images by adjusting its internal weights (i.e., the image data itself) during the training process. It takes the image pixels as parameters and optimizes these parameters through the loss function calculated by the loss network, thus achieving the expected effect of style change.

Loss Network: The research utilizes a loss network that comprises two main elements: style loss, and content loss. The Visual Geometry Group-19 (VGG-19) model is applied to extract these two features from the input as well as the output images, and these features are subsequently used to calculate the loss function. The procedure entails a continuous refinement of the loss values to achieve optimization.

2.2 Details of the Related Methods

(1) Data Preprocessing

For an input jpg image, resize it to a uniform size, and then convert it into a four-dimensional tensor for model input. Next, calculate the Gram matrix for the input style image, which is convenient for later use in the VGG model to compute the style loss.

(2) Gram Matrix

$$\mathrm{G} = \mathrm{A}^{\mathrm{T}}\mathrm{A} = \begin{bmatrix} a_1^T \\ a_2^T \\ \cdot \\ \cdot \\ \cdot \\ a_n^T \end{bmatrix} [a_1\ a_2 \ldots a_n\,] = \begin{bmatrix} a_1^T a_1 & a_1^T a_2 & \ldots & a_1^T a_n \\ a_2^T a_1 & a_2^T a_2 & \ldots & a_2^T a_n \\ & \ldots & & \\ a_n^T a_1 & a_n^T a_1 & \ldots & a_n^T a_n \end{bmatrix} \tag{1}$$

As shown in formula (1), the Gram matrix is the inner product operation of matrices, which involves flattening the image parameters to a one-dimensional vector and performing the inner product after the matrix transpose operation. For example, if the parameter extracted by the input image is [ch, h, w], it can be transformed into matrices of [ch, hw] and [hw, ch], and then their inner product is calculated.

Within style transfer methodologies, the derived feature map encapsulates the characteristics of the image, with each numerical value signifying the strength of a particular feature. The Gram matrix serves as an indicator of the interrelation among these

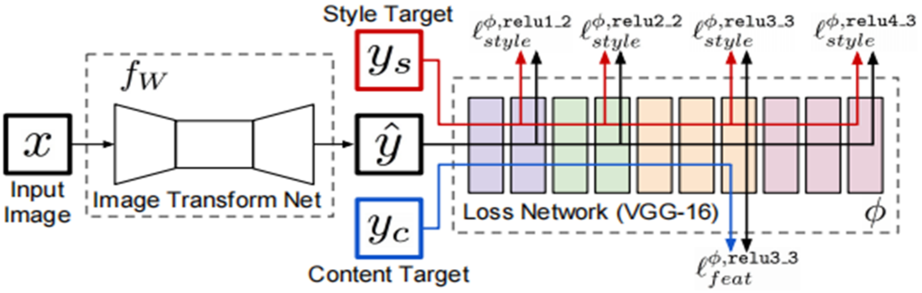

Style Target
$\ell^{\phi,\mathrm{relu1\_2}}_{style}$
$\ell^{\phi,\mathrm{relu2\_2}}_{style}$
$\ell^{\phi,\mathrm{relu3\_3}}_{style}$
$\ell^{\phi,\mathrm{relu4\_3}}_{style}$
y_s
f_W
x
$\hat{y}$
Input Image
Image Transform Net
y_c
Loss Network (VGG-16)
ϕ
$\ell^{\phi,\mathrm{relu3\_3}}_{feat}$
Content Target

Figure 1: Conceptual diagram of the transformation network and loss function extraction network (Johnson, Alahi & Fei-Fei, et al, 2016)

features. By multiplying the values within the same dimension, the features are enhanced, which is why the Gram matrix is employed to encapsulate the style of an image. To ascertain the stylistic discrepancy between the two images, it is sufficient to examine the discrepancies in their respective Gram matrices.

(3) Loss Function Setup

Content Loss Function:

$$l_{feat}^{\Phi,j}(y_{in},y)=\frac{1}{C_jH_jW_j}\|\Phi_j(y_{out})-\Phi_j(y)\|_2^2$$
(2)

Where: y_{out} represents generated image, represents target image; $\Phi_j(y_{out})$ represents the feature matrix extracted from the generated image by VGG, $\Phi_j(y)$ stands for the feature matrix of the target image, while C, H, and W are the parameters defining the image's dimensions, indicating the number of channels, height, and width of the image, respectively.

Style Loss Function:

$$l_{style}^{\Phi,j}(y_{out},y)=\|G_j^{\Phi}(y_{out})-G_j^{\Phi}(y)\|_F^2 \qquad (3)$$

Where: y_{out} represents the output image, y represents the target image; $G_j^{\Phi}(y_{out})$represents the gram matrix of the style image, $G_j^{\Phi}(y)$ represents the gram matrix of the target image.

Full Variational Loss Function:

$$l_{pixel}(y_{out},y)=\|y_{out}-y\|_2^2/CHW \qquad (4)$$

To address the issue of a large number of high-frequency noise points in generated images, the image is smoothed by making the pixels similar to their surrounding adjacent pixels.

(4) VGG-19 network

The VGG network can extract features from input images and output image categories, while image style transfer technology is mainly used to generate images corresponding to the input features and output them, as shown in Figure 4. Through VGGNet, the earlier convolutional layers extract style features from the images, while the later fully connected layers convert these features into category probabilities. The shallow layers extract relatively simple features, such as brightness and dot characteristics; the deeper layers extract more complex features, such as determining whether facial features are present or if a certain specific object appears.

The VGG-19 network architecture comprises a total of 16 convolutional layers alongside 5 pooling layers. This manuscript extracts style features from layers 0, 5, 10, 15, and 20, while content features are derived from the output of layer 16, all of which are employed to compute the loss function.

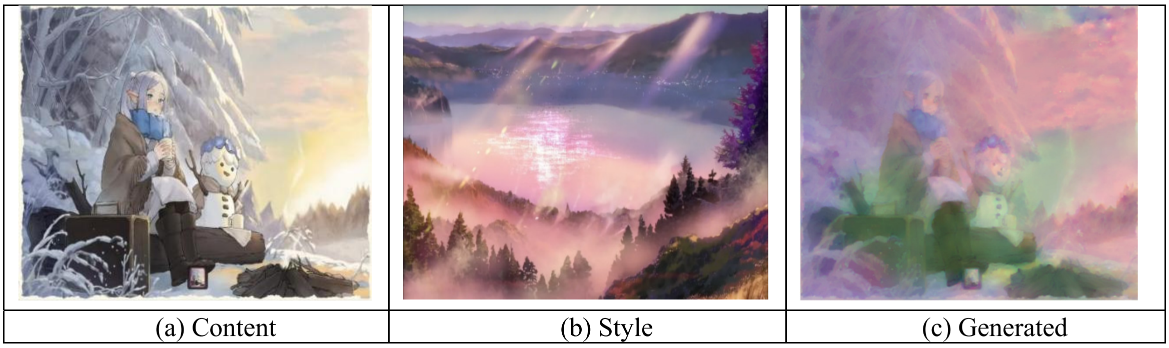
3 RESEARCH RESULT

To assess the performance of the algorithm, the FID and SSIM metrics were utilized, which are standard measures for quantitatively evaluating the quality of synthetic images.

SSIM: SSIM is a measurement that assesses the similarity between two images based on their structural fidelity, luminance, and contrast. The SSIM index ranges from -1 to 1, where a score of 1 indicates that the images are the same.

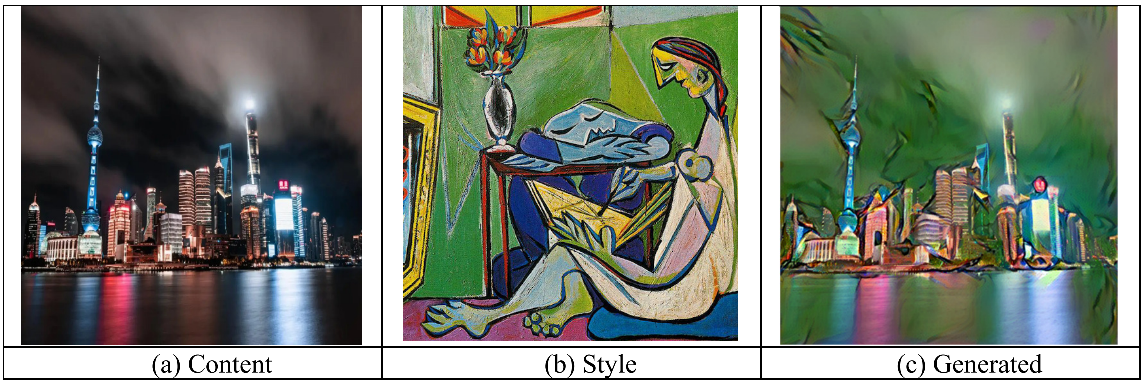
FID: FID is a measurement that gauges the discrepancy between synthetic images and actual images. It relies on the Fréchet distance, a technique for calculating the divergence between two Gaussian distributions. A reduced FID score implies that the artificial images are more statistically similar to the authentic images.

The results are as shown in Figures 2-4, and Table 1.



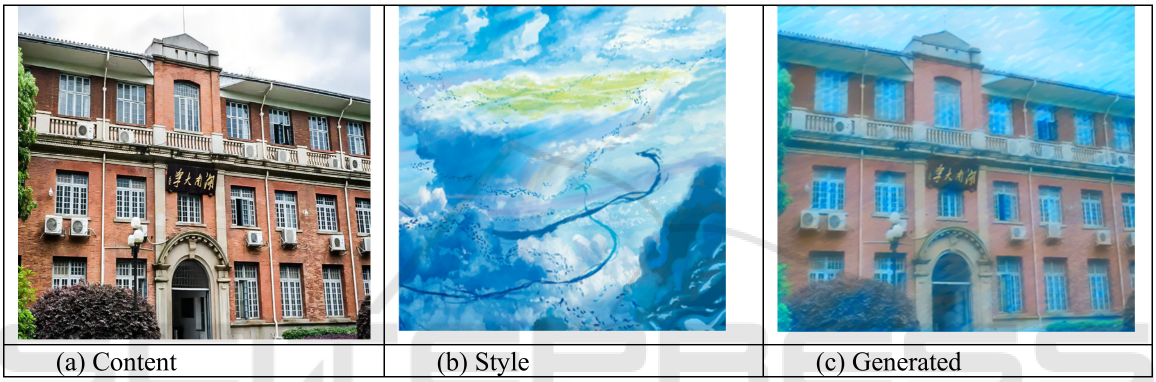
(a) Content (b) Style (c) Generated

Figure 2: A content image (a comic drawing) alongside a style image (a still from a Makoto Shinkai film), along with the resultant composite image (Photo/Picture credit: Original).



(a) Content (b) Style (c) Generated

Figure 3: Contains the content image (Shanghai scenery) and the style image (a Picasso painting) along with the final generated image (Photo/Picture credit: Original).



(a) Content (b) Style (c) Generated

Figure 4: Contains the content image (Hunan University) and the style image (a scene from a Makoto Shinkai film) along with the final generated image (Photo/Picture credit: Original).

Table 1: (Translation): Table of SSIM and FID Metrics for Figures 1 to 3.

	1	2	3
SSIM	0.63	0.59	0.4
FID	1.70	0.67	3.21

4 FUTURE EXPECTATION

Preserving the fidelity of the image content during the process of style transfer poses a significant challenge. It is crucial for style transfer techniques to retain the original content details of the image to the greatest extent possible while altering its stylistic presentation.

Possible solutions: Use more complex and suitable network structures to extract higher-quality content features. For example, research by Jing has shown that graph neural networks have excellent graph data modeling capabilities and the ability to finely grasp complex relationships. Applying them to style transfer is expected to enhance the fineness and accuracy of style conversion (Jing, Mao, Yang, et al, 2022).

Employ regional content preservation techniques: Some style transfer methods allow for the protection of specific areas of the image during style conversion to maintain the accuracy of key content. This can be achieved through the masking techniques or local area optimization proposed by Sun (Sun, Li, Xie, et al, 2022).

Introduction of attention mechanisms: Incorporating attention mechanisms into neural networks can help the model focus more on the key parts of the content, thus better preserving the original content during style transfer. Zhu proposed a self-attention module for learning key style features; and a style inconsistency loss regularization module to promote consistent feature learning and achieve consistent stylization (Zhou, Wu, Zhou, 2023). Wang used the multi-head attention mechanism of Transformer to first identify features in the style image that match the content image and then applied edge detection networks to the content image for refined edge feature extraction (Wang, 2023).

Future possible directions: Domains such as comics or animation: Future style transfer technology may see significant innovation in these animation domains. Wang, facing this potential trend, proposed a method for cartoonizing images (Wang, Yu, 2020). This field has a strong demand for style conversion. For example, people who enjoy this field often like works to be interpreted in their favorite style, and the application of deep learning in style transfer can better learn and understand unique features, especially line style and color selection; improve the preservation of details of different artistic styles to ensure that the converted images still have the recognizability of different artistic styles; combined with the current large models based on Transformer that can generate style images from text, it can greatly enrich the personalized experience of users.

5 CONCLUSION

In today's continuous development of style transfer technology, its application value is constantly being explored, providing a new path for the pursuit of aesthetics. This study focuses on the method of achieving style transfer using CNN technology and mainly adopts the VGG model for exploration. Feeding manipulated images into the VGG model allows for the extraction of style and content characteristics from the images, and then the perceptual loss function can be optimized. Using these optimized features, the style transfer of images is achieved through a feed-forward network, yielding fairly good results, with FID values generally below 5, indicating that actual images and the generated images after style transfer are still very close in content. The future development prospects are then discussed, and it can be found that there are different methods to optimize style transfer technology, such as introducing attention mechanisms to optimize style transfer results or using other network structures. It is expected that the application technology of style transfer will continue to be developed and improved in the future, bringing users richer and more personalized experiences.

REFERENCES

Gatys, L.A., Ecker, A.S. and Bethge, M., 2016. Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, pp. 2414-2423.

Gatys, L., Ecker, A. and Bethge, M., 2016. A neural algorithm of artistic style. Journal of Vision.

Johnson, J., Alahi, A. and Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution.

Jing, Y.C., Mao, Y.N., Yang, Y.D. et al., 2022. Learning graph neural networks for image style transfer. In: European Conference on Computer Vision. Cham: Springer, pp. 111-128.

Liu, J.X. and Zhou, J., 2023. A review of style transfer methods based on deep learning. School of Computer and Information Science / Software College, Southwest University.

Si, Z.Y. and Wang, J.H., 2024. Research on image style transfer method based on improved generative adversarial networks. Journal of Fuyang Teachers College (Natural Science Edition), 41(02), pp. 30-37.

Sun, F.W., Li, C.Y., Xie, Y.Q., Li, Z.B., Yang, C.D. and Qi, J., 2022. A review of deep learning applications in occluded target detection algorithms. Computer Science and Exploration, 16(06), pp. 1243-1259.

Wang, X.R. and Yu, J.Z., 2020. Learning to cartoonize using white-box cartoon representations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, pp. 8087-8096.

Wang, Z.P., 2023. Research on image style transfer algorithms based on deep learning. East China Normal University.

Zhou, Z., Wu, Y. and Zhou, Y., 2023. Consistent arbitrary style transfer using consistency training and self-attention module. IEEE Transactions on Neural Networks and Learning Systems, [Epub ahead of print].

Zhu, J.Y., Park, T., Isola, P. et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, pp. 2242-2251.

Zhang, H.H. and Huang, L.J., 2023. Efficient style transfer based on convolutional neural networks. Guangdong University of Business and Technology Art College, Zhaoqing, Guangdong 526000; Computer College.